Inference with Transposable Data: Modeling the Effects of Row and Column Correlations

Genevera I. Allen
&
Robert Tibshirani

April 2, 2010

Abstract

We consider the problem of large-scale inference on the row or column variables of data in the form of a matrix. Often this data is *transposable*, meaning that both the row variables and column variables are of potential interest. An example of this scenario is detecting significant genes in microarrays when the samples or arrays may be dependent due to underlying relationships. We study the effect of both row and column correlations on commonly used test-statistics, null distributions, and multiple testing procedures, by explicitly modeling the covariances with the matrix-variate normal distribution. Using this model, we give both theoretical and simulation results revealing the problems associated with using standard statistical methodology on transposable data. We solve these problems by estimating the row and column covariances simultaneously, with transposable regularized covariance models, and de-correlating or *sphering* the data as a pre-processing step. Under reasonable assumptions, our method gives test statistics that follow the scaled theoretical null distribution and are approximately independent. Simulations based on various models with structured and observed covariances from real microarray data reveal that our method offers substantial improvements in two areas: 1) increased statistical power and 2) correct estimation of false discovery rates.

Keywords: *multiple testing, false discovery rate, transposable regularized covariance models, large-scale inference, covariance estimation, matrix-variate normal, empirical null*

1 Introduction

As statisticians, we often make assumptions when constructing a model to ease computations or employ existing methodologies. When analyzing matrix data, we often assume that the variables along one dimension (say the columns) are independent, allowing us to pool these observations to make inferences on the variables along the other dimension (rows). In microarrays, for example, it is common to assume that the arrays are independent observations when computing test statistics, allowing us to assess differential expression in genes.

Since we are testing many row variables (for example, genes) simultaneously, we commonly correct for multiple testing using procedures that theoretically are known only to control error measures when the row variables are independent or follow limited dependence structures. Thus, for inference with matrix data, we often make assumptions of independence or limited dependencies among the row variables and among the column variables to be able to employ existing statistical methodologies. What if these assumptions are incorrect? What if this matrix data is in fact *transposable*, meaning that potentially both the rows and/or columns are correlated?

In this paper, we consider the problem of testing the significance of row variables in a data matrix where there are correlations among the rows, or among the columns, or among both. We study the behavior of standard statistical methodology on transposable data and then propose a method to directly account for the dependencies when conducting inference.

Throughout this paper, we often refer to the example of detecting genes that are differentially expressed between two classes in microarray data. These genomic datasets contain complicated correlation structures. Genes in similar pathways, for example, are usually highly positively correlated. Other genes may encode proteins that act as inhibitors leading to negative correlations. In the analysis of microarrays, it is common to assume that the arrays are independent. Many have suggested, however, that this may not be correct (Efron, 2009; Leek and Storey, 2008; Owen, 2005; Qiu et al., 2005), due to the measurement process or latent variables. Arranged in the form of a matrix, this means that both the row (gene) and column (array) variables could be dependent, indicating that the data could be *transposable*.

While we focus on the example of detecting significant genes in the two-class microarray, our methods can be applied to many examples of large-scale inference with transposable data. These include: testing the significance of proteins, genes, or isoforms in data such as protein arrays and next-generation sequencing data, testing the significance of voxels in functional magnetic resonance imaging data, and testing the significance of biomarkers in three-way data where measurements are taken on multiple subjects at several time points or in many different laboratories. In all of these examples, the assumptions of independence along one dimension of the data is questionable.

We begin by introducing two examples that we will refer to throughout this paper. The first is a two-class microarray study of cardiovascular disease (Efron, 2009). We will refer to this as the "Cardio" data. This data has $m = 20,426$ genes and $n = 63$ arrays consisting of 44 controls and 19 diseased patients. The second is a two-class microarray study of two types of Leukemia cancer (Golub et al., 1999), which we will refer to as the "Leukemia" data. This data has $m = 3,701$ filtered genes and $n = 72$ arrays with 25 and 47 samples in each subtype. For each of these datasets, we calculate the two-sample t-statistic for each gene and compare their distribution to that of the theoretical null distribution in Figure 1. We see that the t-statistics are over-dispersed compared to their theoretical null distributions. This could be due to the highly correlated nature of the thousands of genes, or another cause could be correlations among the arrays. In fact, the permutation tests of Efron (2009) reject the null hypothesis of independent arrays for both of these microarrays.

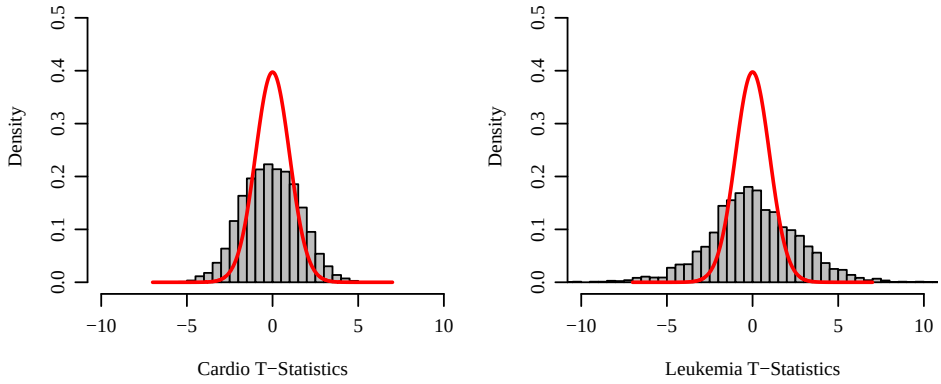

Figure 1: *Histograms of two-sample t-statistics for the "Cardio" data (left) and the "Leukemia" data (right). Log intensity values were used with the genes and arrays centered. The theoretical null distribution, the t-distribution with 70 degrees of freedom (left) and 61 degrees of freedom (right), is drawn in red.*

When studying inference with transposable data, the effects of row and column correlations must be considered separately. Since the columns are generally considered to be independent, population column correlations lead to the used of incorrect test statistics and null distributions which in turn result in problems when correcting for multiple testing. Row correlations lead to the much discussed problem of multiple testing dependence (Benjamini and Yekutieli, 2001; Hommel, 1986; Leek and Storey, 2008; Sarkar, 2008; Storey et al., 2004).

We propose to study and solve these problems by modeling row and column correlations using the mean-restricted matrix-variate normal distribution (Allen and Tibshirani, 2010) described in Section 2. The first half of our paper is devoted to studying the effects of these correlations on test statistics and their theoretical null distributions, Section 2.2, and on power and multiple testing procedures through a simulation study in Section 3.2. Interestingly, this study finds the following results.

1. Unanticipated column correlations dramatically alter the null distributions of test statistics leading to the use of incorrect test statistics, null distributions and estimates of the FDR.

2. Row correlations do not seem to affect the estimates of the FDR.

The later half of our paper is focused on solving the problems associated with row and column correlations by directly making use of the correlation structure. In Section 4, we simultaneously estimate row and column covariances using transposable regularized covariance models (Allen and Tibshirani, 2010). We then present an algorithm to *sphere* or de-correlate the rows and columns so that they are approximately independent. This algorithm is to be used as a pre-processing step and in conjunction with standard multiple

testing procedures. Simulation results using our sphering algorithm are presented in Section 5 under various models on both structured covariance and real microarray covariance examples. These reveal two important results:

(c) Sphering can alter the rank of the test statistics leading to an ordering with higher statistical power.

(d) Sphering often leads to substantial improvements in the estimation of the FDR.

We conclude with a discussion of our study and methods in Section 6.

2 Theoretical Framework

In this section, we first present a matrix decomposition model based on the mean-restricted matrix-variate normal in Section 2.1. Then, going back to the two-class microarray example, we consider the test statistic for a single gene. Since the arrays are usually assumed to be independent, the two-sample z and t-tests are used commonly to assess differential expression. We give the theoretical null distributions for these test statistics under our model with column correlations in Section 2.2.

2.1 Model

We propose to study row and column correlations through a simple matrix decomposition model based on the matrix-variate normal. We motivate the use of this distribution through the example of microarrays.

In microarray data, the genes are often assumed to follow a multivariate normal distribution with the arrays independent and identically distributed. Since we aim to study the effects of array correlations, we need a parametric model that has the flexibility to model either array independence or various array correlation structures. To this end, we turn to the mean-restricted matrix-variate normal introduced in Allen and Tibshirani (2010). (We also note that Efron (2009) proposes the matrix-variate normal as a model for microarrays). This distribution, denoted as $\mathbf{X} \sim N_{m,n}(\nu, \mu, \mathbf{\Sigma}, \mathbf{\Delta})$, has separate mean and covariance parameters for the rows, $\nu \in \Re^m$ and $\mathbf{\Sigma} \in \Re^{m \times m}$, and columns, $\mu \in \Re^n$ and $\mathbf{\Delta} \in \Re^{n \times n}$. Thus, we can model array correlations directly though the covariance matrix $\mathbf{\Delta}$. If the matrix is transformed into a vector of length np, we have that $\text{vec}(\mathbf{X}) \sim N(\text{vec}(\mathbf{M}), \mathbf{\Omega})$, where $\mathbf{M} = \nu \mathbf{1}_{(n)}^T + \mathbf{1}_{(m)}$ and $\mathbf{\Omega} = \mathbf{\Delta} \otimes \mathbf{\Sigma}$. Also, the commonly used multivariate normal is a special case of the distribution. If $\mathbf{\Delta} = \mathbf{I}$ and $\mu = \mathbf{0}$, then $\mathbf{X} \sim N(\nu, \mathbf{\Sigma})$. In fact, all marginal models of the matrix-variate normal are multivariate normal, meaning that both the genes and arrays separately are multivariate normal. Further properties of this distribution are given in Allen and Tibshirani (2010).

In our matrix decomposition model, we will assume that the data, $\mathbf{X}$, has m rows and n columns. We define the overall row means as $\nu \in \Re^m$ and column means as $\mu \in \Re^n$. The

covariance of the rows is $\mathbf{\Sigma} \in \Re^{m \times m}$ and the covariance of the columns is $\mathbf{\Delta} \in \Re^{n \times n}$. Then, we decompose the data into a mean, signal, and correlated noise matrix as follows.

$$\mathbf{X}_{m\times n} = \mathbf{M}_{m\times n} + \mathbf{S}_{m\times n} + \mathbf{N}_{m\times n}, \tag{1}$$

$$\begin{aligned} \text{where } \mathbf{M} &= \nu \mathbf{1}_{(n)}^T + \mathbf{1}_{(m)} \mu^T \text{ (mean matrix)}, \\ \mathbf{S} &\text{ is problem specific (signal matrix)}, \\ \mathbf{N} &\sim N_{m,n}(\mathbf{0}, \mathbf{0}, \mathbf{\Sigma}, \mathbf{\Delta}) \text{ (noise matrix)}. \end{aligned}$$

Thus, $\mathbf{X} - \mathbf{S} \sim N_{m,n}(\nu, \mu, \mathbf{\Sigma}, \mathbf{\Delta})$, meaning that after removing the signal, the data follows a mean-restricted matrix-variate normal distribution.

For the example of the two-class microarray, we let there be n_1 arrays in class one, with indices denoted by $\mathcal{C}_1$, and n_2 in class two, $\mathcal{C}_2$. (For simplicity of notation, we assume that the first n_1 arrays are in class one and the last n_2 arrays are in class two.) The class signals, or the gene means for each class are defined as $\psi_1 \in \Re^m$ and $\psi_2 \in \Re^m$. Then, the signal matrix, $\mathbf{S}$, can be written as follows.

$$\mathbf{S} = \left[\psi_1 \mathbf{1}_{(n_1)}^T \;\; \psi_2 \mathbf{1}_{(n_2)}^T \right].$$

There are several remarks to make regarding this model. First, prior to analyzing data, it is common to standardize the rows. Sometimes this two-way data is *doubly-standardized*, or both the rows and columns are iteratively scaled (Efron, 2009; Olshen and Rajaratnam, 2010). Here, we center both the rows and columns through the mean matrix $\mathbf{M}$, but do not directly scale them. Instead, we allow the diagonals of the covariance matrices of the rows, $\mathbf{\Sigma}$, and columns $\mathbf{\Delta}$, to capture the differences in variablities. Thus, our model keeps the mean and variances separate in the estimation process.

2.2 Null Distributions: The Two-Class Problem

In this section, we study the effect of column correlations on the theoretical null distribution of two-sample test statistics computed for a single row of the data matrix. More specifically, we calculate the distributions of test statistics under our matrix decomposition model instead of the typical two-sample framework where samples are drawn independently from two populations. This corresponds to considering a single test for differential expression of gene i between the two classes.

In the familiar two-sample hypothesis testing problem, we have a vector $\mathbf{x} = [\mathbf{x}_1 \;\; \mathbf{x}_2]$ with $\mathbf{x}_1$ of length n_1 and $\mathbf{x}_2$ of length n_2 where the elements of each vector are $x_{1,i} \overset{iid}{\sim} N(\psi_1, \sigma^2)$ and $x_{2,i} \overset{iid}{\sim} N(\psi_2, \sigma^2)$. We wish to test whether there is a shift in means between the two classes, namely

$$H_0: \;\; \psi_1 = \psi_2 \;\text{ vs. }\; H_1: \;\; \psi_1 \neq \psi_2. \tag{2}$$

Throughout this paper, we will assume that the variances σ^2 are equal between the two classes, a common assumption in microarrays.

If the variance, σ^2 is known, we have the familiar two-sample Z-statistic,

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{c_n}}, \text{ with } Z \sim N\left(\frac{\psi_1 - \psi_2}{\sigma\sqrt{c_n}}, 1\right),$$

where $\bar{x}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_i$ and $c_n = \frac{1}{n_1} + \frac{1}{n_2}$.

Now, going back to our matrix decomposition model, we wish to know the distribution of the Z-statistic for each row when there are column correlations.

Theorem 1 *Let* $\mathbf{x} = [\mathbf{x}_1 \;\; \mathbf{x}_2] \sim N_{1,n}\left(0, [\psi_1\mathbf{1}_{(n_1)} \;\; \psi_2\mathbf{1}_{(n_2)}], \sigma^2, \mathbf{\Delta}\right)$. *Then,*

$$Z \sim N\left(\frac{\psi_1 - \psi_2}{\sigma\sqrt{c_n}}, \frac{\eta}{c_n}\right) \tag{3}$$

where $\eta \triangleq \sum_{j=1}^{n}\left(\frac{1}{n_1}\sum_{i\in\mathcal{C}_1} L_{ij} - \frac{1}{n_2}\sum_{i\in\mathcal{C}_2} L_{ij}\right)^2$ *and* $\mathbf{L}$ *is the matrix square root of* $\mathbf{\Delta}$.

In terms of the decomposition (1), the assumptions of Theorem 1 correspond to a row vector previously centered by ν and μ, with signal $[\psi_1\mathbf{1}_{(n_1)} \;\; \psi_2\mathbf{1}_{(n_2)}]$, column covariance $\mathbf{\Delta}$, and row variance σ^2, the diagonal element of $\mathbf{\Sigma}$. For microarrays, the result states that when the columns (arrays) are correlated, the variance of the Z-statistic is inflated or deflated by η, a function of the column covariance. Notice that if $\mathbf{\Delta} = \mathbf{I}$, $\eta = c_n$ and the variance of Z is one. If there is only column correlation within the two classes we have the following result.

Corollary 1 *Assume* $\mathbf{x} = [\mathbf{x}_1 \;\; \mathbf{x}_2]$ *with* $\mathbf{x}_1 \sim N_{1,n_1}(0, \psi_1\mathbf{1}_{(n_1)}, \sigma^2, \mathbf{\Delta}_1)$ *and* $\mathbf{x}_2 \sim N_{1,n_2}(0, \psi_2\mathbf{1}_{(n_2)}, \sigma^2, \mathbf{\Delta}_2)$ *such that* $\mathrm{Cov}(\mathbf{x}) = \mathbf{\Delta} = \begin{pmatrix} \mathbf{\Delta}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Delta}_2 \end{pmatrix}$, *then*

$$Z \sim N\left(\frac{\psi_1 - \psi_2}{\sigma\sqrt{c_n}}, \frac{\eta_1 + \eta_2}{c_n}\right) \tag{4}$$

where $\eta_k \triangleq \frac{1}{n_k^2}\sum_{i=1}^{n_k}\left(\sum_{j=1}^{n_k} L_{k,ij}\right)^2 = \frac{1}{n_k^2}\sum_{i=1}^{n_k}\sum_{j=1}^{n_k}\mathbf{\Delta}_{k,ij}$ *for* $k = 1, 2$ *and* $\mathbf{L}_k$ *is the matrix square root of* $\mathbf{\Delta}_k$.

In both of the previous results, we assumed that the row variance, σ^2 was known. However, in most microarray experiments this is not known and must be estimated. With σ^2 unknown, for testing the hypothesis (2), the two-sample t-statistic is used.

$$T = \frac{\bar{x}_1 - \bar{x}_2}{s_{\mathbf{x}_1,\mathbf{x}_2}\sqrt{c_n}}, \qquad s^2_{\mathbf{x}_1,\mathbf{x}_2} = \frac{\sum_{i\in y\mathcal{C}_1}(x_{1,i} - \bar{x}_1)^2 + \sum_{i\in\mathcal{C}_2}(x_{2,i} - \bar{x}_2)^2}{n_1 + n_2 - 2} \tag{5}$$

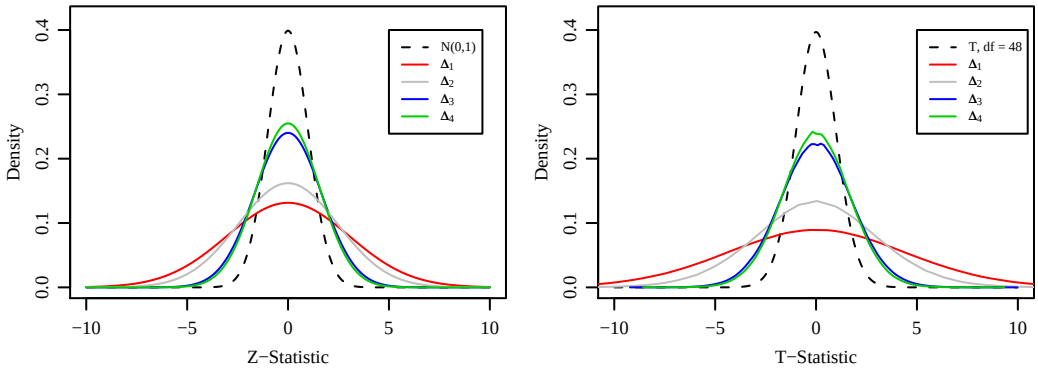

Figure 2: *Comparison of theoretical null distributions for the two-sample Z-statistic (left) and T-statistic (right) under various column covariance scenarios given in Section 2.2. Variances of the Z-statistics were calculated by the result in Theorem 1, while the densities of the T-statistics were estimated via a simulation with one million replicates.*

with c_n and $\bar{x}_k$ as previously defined. Under the null hypothesis, $T \sim t_{(n-2)}$, while under the alternative, $T \sim t(\delta)_{(n-2)}$, a non-central t distribution with non-centrality parameter $\delta = (\psi_1 - \psi_2)/(\sigma\sqrt{c_n})$.

When there are column correlations as in the assumptions of Proposition 1, however, the distribution of T does not have a closed form. (The square of the pooled sample standard deviation is no longer distributed as a Chi-squared random variable and the numerator and denominator of T are not independent.) Hence, we explore the effects of column correlations on the T-statistic through a small simulation study. Data is simulated according to the assumptions of Theorem 1 with $n = 50$ columns with $n_1 = n_2 = 25$ in each class. Four structured covariance matrices were used to assess the Z and T-statistics under the column correlations scenarios, as given below.

- $\mathbf{\Delta}_1 : \mathbf{\Delta}_{1,ij} = 0.9^{|i-j|}$.
- $\mathbf{\Delta}_2$: Blocked diagonal with blocks of size 10. Within each block, $\mathbf{\Delta}_{2,ij} = 0.9^{|i-j|}$.
- $\mathbf{\Delta}_3 : \mathbf{\Delta}_{3,ij} = 0.5^{|i-j|}$.
- $\mathbf{\Delta}_4$: Blocked diagonal with blocks of size 10. Within each block, $\mathbf{\Delta}_{4,ij} = 0.5^{|i-j|}$.

Figure 2 reveals the effect of column correlations on the distributions of Z and T. We see that column correlations can cause dramatic over-dispersion of the test statistics compared to their theoretical null distribution. This is a possible explanation to the over-dispersion seen in the real microarray examples of Figure 1. Compared to the variance of the Z-statistic, the T-statistic appears to be even more affected by column correlations. This is confirmed in Table 1 where we present the variances of the Z-statistic calculated by Theorem 1 and the variances of the T-statistic estimated by Monte Carlo simulation. Indeed, small amounts of correlation in the columns can cause a dramatic increase in the variance of the T-statistic.

In this section, we have shown how the distribution of T and Z-statistics behave when columns or arrays are correlated. When analyzing microarrays, however, many have advocated using a non-parametric method, estimating the null distribution by permutations Ge

	Var(Z-statistic)	Var(T-statistic)
$\mathbf{\Delta}_1$	9.215	19.94 (0.029)
$\mathbf{\Delta}_2$	6.069	9.492 (0.0144)
$\mathbf{\Delta}_3$	2.76	3.197 (0.00472)
$\mathbf{\Delta}_4$	2.45	2.79 (0.00411)

Table 1: *Variances of the two-sample Z and T-statistics under various column correlation scenarios as given in Section 2.2. Variances of the T-statistics were estimated via simulation with one million replicates. The theoretical variance of the Z-statistic should be one, and 1.022 for the T-statistic.*

et al. (2003); Storey and Tibshirani (2003); Tusher et al. (2001). For the two-class microarray, one would permute the class labels and calculate the T-statistic for each permutation. These permutations form a null distribution, as under the null hypothesis (2), the class means are the same. Thus, each permutation of the labels is equally likely. When the arrays are correlated, however, this assumptions fails. Each permutation of the columns is not equally likely under the null due to the array covariance structure. While we do not explore the behavior of the permutation nulls further in this section, we include permutation-based methods from Storey and Tibshirani (2003) in our simulation study in the following section.

3 Study: Dependence and Multiple Testing

In the previous section, we presented the theoretical null distributions of commonly used test statistics for a single two-sample test statistic when the columns are correlated. With transposable data, however, one needs to test possibly tens of thousands of row variables, thus creating a problem of multiplicity. In this section, we first review some multiple testing procedures that are known to control errors under certain types of dependencies. We then present a series of simulations to study the behavior of commonly used multiple testing procedures when the rows and columns are correlated.

3.1 Background

A common error measure for controling the number of false positives in microarrays is the False Discovery Rate (FDR). This is the expectation of the False Discovery Proportion (FDP): let V be the number of false positives and R be the total number of rejections, then $q = FDR = \mathrm{E}(V/R|R > 0)$. Typically, investigators seek to control the FDR at $q = 0.1$, meaning that on average 10% of rejections are false.

The step-up method of Benjamini and Hochberg (1995) is one of the most widely used methods for controling the FDR. Benjamini and Yekutieli (2001) have shown that this method controls the FDR under types of positive dependence, specifically *positive regression dependence*, and Sarkar (2008) has relaxed this assumption to slightly broader forms of positive dependence. This may not be appropriate for all types of transposable data,

especially microarrays where we expect some negative correlations between genes. Alternatively, Benjamini and Yekutieli (2001) have shown that dividing the thresholds in the step-up procedure by a constant controls the FDR under arbitrary dependencies.

Another commonly used method to control the FDR is based on re-sampling or permutation distributions (Ge et al., 2003; Storey, 2002; Tusher et al., 2001; Yekutieli and Benjamini, 1999). Theoretically, these methods are only known to control the FDR asymptotically under types of *weak dependence*, which encompasses forms of local dependence such as finite blocks (Storey et al., 2004). Thus, there could be many transposable data sets in which the row variables do not satisfy these dependence structures. (We also note that applying the step-up method to the permutation-adjusted p-values is equivalent to the direct FDR estimation via re-sampling (Storey et al., 2004)).

Also, to directly account for correlations, Efron (2004, 2007) proposed a method to fit an *empirical null* to the data. One can then estimate the local FDR and then the FDR by averaging the local FDR over the tail regions.

3.2 Simulation Study

We study the effects of both row and column correlations on standard statistical methodology used for large-scale inference through a simulation study based on our matrix decomposition model. We compare FDR estimates of four types of FDR-controlling procedures to the true false discovery proportion (FDP). The four methods we compare are the step-up method of Benjamini and Hochberg (1995), the step-up method for control under arbitrary dependence of Benjamini and Yekutieli (2001), the permutation-based method of Storey and Tibshirani (2003), and the method based on the empirical null and local FDR's of Efron (2007). The two-sample t-statistic was used for all methods with p-values computed by comparing it to the $t_{(n-2)}$ distribution for the step-up procedures. We used 1000 permutations for the permutation-based method. The defaults in the localfdr package available on CRAN, the R language repository. These defaults fit the null distribution as a natural spline with seven degrees of freedom, for the empirical null-based method.

Our simulation study is structured as follows. The data is simulated under the matrix decomposition model (1) and is of size 250 by 50. The first 50 rows are non-null with a two-class signal matrix given by $\psi_{1,1:25} = 0.5$, $\psi_{1,26:50} = -0.5$, $\psi_{2,1:25} = -0.5$, $\psi_{2,26:50} = 0.5$ and the last 200 elements of ψ_1 and ψ_2 equal to zero. We consider two types of row covariances, $\mathbf{\Sigma}_1$ with all positive correlations satisfying the positive regression dependence assumption of (Benjamini and Yekutieli, 2001), and $\mathbf{\Sigma}_2$ with both positive and negative correlations. Both of these row covariances are block diagonal. We simulate data under three column covariances, with the first being the identity, or no column correlations. The others, $\mathbf{\Delta}_1$ and $\mathbf{\Delta}_2$ reflect a local and a class effect, respectively. These simulation covariances are summarized below.

- $\mathbf{\Sigma}_1$: Blocked diagonal with blocks of size 10. Within each block, $\mathbf{\Sigma}_{1,ij} = 0.9^{|i-j|}$.
- $\mathbf{\Sigma}_2$: Blocked diagonal with blocks of size 10. Within each block, $\mathbf{\Sigma}_{2,ij} = (-0.9)^{|i-j|}$.
- $\mathbf{\Delta}_1$: Blocked diagonal with blocks of size 10. Within each block, $\mathbf{\Delta}_{1,ij} = 0.5^{|i-j|}$.
- $\mathbf{\Delta}_2$: Blocked diagonal with blocks of size 25. Within each block, $\mathbf{\Delta}_{2,ij} = 0.5^{|i-j|}$.

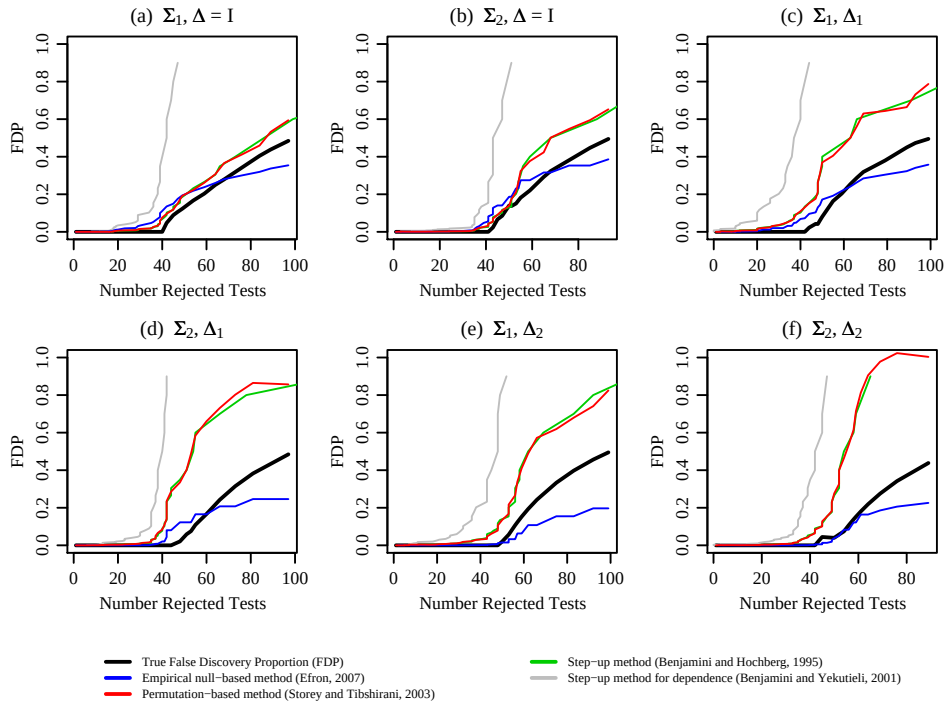

Figure 3: *Simulation Study FDR Curves: The true and estimated false discovery proportions plotted against the number of tests rejected for each of the six simulations. All data was simulated under the matrix decomposition model,* (1)*, with parameters given in Section 3.2.*

We present plots of the true FDP verses the number of hypotheses rejected for the four methods for one realization of each of the six simulation scenarios in Figure 3. We note that the lines above the true FDP curve denote conservative FDR estimates. In Table 2, we report the true and estimated false discovery proportions (FDP) when fixed numbers of hypotheses are rejected (40, 45, 50, 55 and 60 tests). Results are averaged over ten simulations with the standard error also reported.

This simulation study reveals several interesting results. First, dependencies among the rows do not seem to effect FDR estimation with the four multiple testing procedures. When, $\mathbf{\Delta} = \mathbf{I}$ as in simulations (a) and (b), the methods generally conservatively estimate the true FDP. This is noteworthy since besides the method of Benjamini and Yekutieli (2001), there are limited theoretical results supporting FDR control under various dependencies.

When there are even moderate correlations between columns, simulations (c) through (f), the four methods give poor estimates of the FDR. The step-up method and the permutation-based method perform similarly. They both give extremely conservative estimates of the FDP when there is either a local or a class effect among the columns. Thus, when using these methods for controlling the FDR at $q = 0.1$, for example, one would reject

	True FDP	(Benjamini & Hochberg, 1995)	(Benjamini & Yekutieli, 2001)	(Storey & Tibshirani, 2003)	(Efron, 2007)
(a) $\mathbf{\Sigma}_1, \mathbf{\Delta} = \mathbf{I}$					
40 tests	0.0725 (0.033)	0.0723 (0.022)	0.387 (0.091)	0.0742 (0.022)	0.257 (0.068)
45 tests	0.104 (0.034)	0.103 (0.026)	0.545 (0.11)	0.105 (0.026)	0.286 (0.07)
50 tests	0.124 (0.033)	0.147 (0.028)	0.715 (0.1)	0.149 (0.029)	0.32 (0.069)
55 tests	0.169 (0.025)	0.19 (0.03)	0.823 (0.092)	0.193 (0.03)	0.344 (0.066)
60 tests	0.222 (0.02)	0.255 (0.036)	0.891 (0.069)	0.258 (0.036)	0.376 (0.063)
(b) $\mathbf{\Sigma}_2, \mathbf{\Delta} = \mathbf{I}$					
40 tests	0.035 (0.017)	0.0498 (0.0081)	0.304 (0.049)	0.0513 (0.0082)	0.161 (0.034)
45 tests	0.0711 (0.019)	0.0839 (0.012)	0.512 (0.076)	0.0856 (0.012)	0.209 (0.041)
50 tests	0.12 (0.018)	0.141 (0.021)	0.771 (0.099)	0.143 (0.021)	0.249 (0.045)
55 tests	0.167 (0.016)	0.191 (0.029)	0.822 (0.094)	0.192 (0.029)	0.278 (0.04)
60 tests	0.217 (0.014)	0.243 (0.035)	0.867 (0.074)	0.246 (0.035)	0.311 (0.041)
(c) $\mathbf{\Sigma}_1, \mathbf{\Delta}_1$					
40 tests	0.0075 (0.0053)	0.0588 (0.019)	0.329 (0.087)	0.0578 (0.019)	0.0152 (0.0053)
45 tests	0.0222 (0.0099)	0.134 (0.037)	0.599 (0.11)	0.133 (0.038)	0.0511 (0.025)
50 tests	0.056 (0.016)	0.245 (0.052)	0.847 (0.087)	0.247 (0.053)	0.0858 (0.032)
55 tests	0.111 (0.011)	0.426 (0.04)	1 (0)	0.43 (0.041)	0.139 (0.032)
60 tests	0.178 (0.0083)	0.579 (0.052)	1 (0)	0.585 (0.053)	0.167 (0.028)
(d) $\mathbf{\Sigma}_2, \mathbf{\Delta}_1$					
40 tests	0.005 (0.005)	0.0455 (0.0065)	0.277 (0.04)	0.0439 (0.0065)	0.00846 (0.0041)
45 tests	0.0111 (0.005)	0.111 (0.027)	0.58 (0.09)	0.11 (0.026)	0.0198 (0.0076)
50 tests	0.042 (0.0081)	0.225 (0.046)	0.869 (0.082)	0.225 (0.047)	0.0493 (0.014)
55 tests	0.109 (0.0086)	0.404 (0.034)	1 (0)	0.409 (0.034)	0.0923 (0.017)
60 tests	0.178 (0.0056)	0.552 (0.048)	1 (0)	0.554 (0.048)	0.133 (0.018)
(e) $\mathbf{\Sigma}_1, \mathbf{\Delta}_2$					
40 tests	0.0125 (0.0077)	0.0831 (0.018)	0.476 (0.09)	0.0783 (0.019)	0.0749 (0.024)
45 tests	0.0333 (0.015)	0.164 (0.031)	0.746 (0.097)	0.16 (0.032)	0.117 (0.032)
50 tests	0.078 (0.015)	0.281 (0.027)	0.969 (0.031)	0.276 (0.028)	0.165 (0.032)
55 tests	0.135 (0.012)	0.368 (0.028)	1 (0)	0.364 (0.028)	0.194 (0.028)
60 tests	0.198 (0.011)	0.461 (0.044)	1 (0)	0.458 (0.045)	0.234 (0.028)
(f) $\mathbf{\Sigma}_2, \mathbf{\Delta}_2$					
40 tests	0.0075 (0.0053)	0.0712 (0.016)	0.407 (0.073)	0.066 (0.015)	0.0444 (0.012)
45 tests	0.0311 (0.011)	0.169 (0.022)	0.855 (0.072)	0.163 (0.021)	0.087 (0.019)
50 tests	0.078 (0.012)	0.277 (0.025)	0.99 (0.0095)	0.271 (0.025)	0.132 (0.021)
55 tests	0.144 (0.013)	0.388 (0.037)	1 (0)	0.383 (0.037)	0.167 (0.019)
60 tests	0.198 (0.013)	0.47 (0.044)	1 (0)	0.468 (0.043)	0.197 (0.02)

Table 2: *Simulation Study: The effect of row and column correlations on estimation of the false discovery rates. The true false discovery proportion (FDP) and estimates with standard errors using the step-up, step-up for dependence, permutation, and empirical null based methods, as described in Section 3.2, are given when a pre-specified number of tests are rejected. All simulations were done using the matrix decomposition model,* (1)*, with parameters given in Section 3.2, and repeated ten times.*

less than 45 genes, when in reality one should be permitted to reject around 55 genes. We also see that while the method of Benjamini and Yekutieli (2001) controls the FDR are arbitrary dependencies, in practice this method is much too conservative for general use. On the other hand, the empirical null-based method of Efron (2007) performs inconsistently.

Overall, the results of this simulation study reveal that dependencies among rows do not seem to effect the performance of the multiple testing procedures. On the other hand, the theoretical results of Section 2.2 are confirmed: dependencies among the columns are extremely problematic when conducting large-scale inference.

4 De-Correlating a Matrix

In the previous sections, we have presented theoretical and simulation results demonstrating some of the problems with using standard statistical methodology for making inferences on transposable data. In the remainder of this paper, we present a solution to these problems by directly estimating the covariances and using these to *sphere* or de-correlate the data.

The key to our method, based on the matrix decomposition model (1), is the simultaneous estimation of the row and column covariances. This is important because of the close relationship between the observed row and column covariances. Take, for example, the empirical covariances of a centered data matrix $\mathbf{X}$, $\hat{\mathbf{\Sigma}} = \mathbf{X}\,\mathbf{X}^T/m$ and $\hat{\mathbf{\Delta}} = \mathbf{X}^T\,\mathbf{X}/n$. If we take the singular value decomposition, $\mathbf{X} = \mathbf{U}\,\mathbf{D}\,\mathbf{V}^T$, then $\hat{\mathbf{\Sigma}} = \mathbf{U}\,\mathbf{D}\,\mathbf{U}^T$ and $\hat{\mathbf{\Delta}} = \mathbf{V}\,\mathbf{D}\,\mathbf{V}^T$, i.e. the two covariances estimates share the same eigenvalues. In fact, Efron (2009) shows that the variance of the two correlation matrices is the same. Because of this, population correlations among the rows, for example, often make the columns seem correlated. Thus, estimating the column covariance without accounting for the covariances of the rows is problematic. With Transposable Regularized Covariance Models, we can estimate both $\mathbf{\Sigma}$ and $\mathbf{\Delta}$ simultaneously according the the matrix-variate normal framework. We review these models and discuss their relevance for the example of microarrays in the next section.

4.1 Review: Transposable Regularized Covariance Models

The Transposable Regularized Covariance Model (TRCM) allows us to estimate a non-singular row and column covariance matrix by maximizing a penalized log-likelihood of the matrix-variate normal distribution (Allen and Tibshirani, 2010). The model places a strictly convex penalty on the inverse covariances, or concentration matrices, of the rows and columns. For estimating the covariances in this context, we propose to use a sparsity-inducing penalty, an L_1 penalty, on the concentration matrices. Following from the matrix decomposition model, (1), if we let $\mathbf{N}$ be the noise matrix remaining after removing the means and the signal in the data, then the penalized log-likelihood is as follows.

$$\ell(\mathbf{\Sigma}, \mathbf{\Delta}) = \frac{n}{2}\log|\,\mathbf{\Sigma}^{-1}\,| + \frac{m}{2}\log|\,\mathbf{\Delta}^{-1}\,| - \frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\,\mathbf{N}\,\mathbf{\Delta}^{-1}\,\mathbf{N}^T\right) - \lambda m||\,\mathbf{\Sigma}^{-1}\,||_1 - \lambda n||\,\mathbf{\Delta}^{-1}\,||_1 \tag{6}$$

where $||\,\mathbf{\Delta}^{-1}\,||_1 = \sum_{i=1}^n \sum_{j=1}^n |\,\mathbf{\Delta}^{-1}{}_{ij}\,|$ and λ is a penalty parameter that must be estimated.

We motivate the use of (6) first by discussing practical considerations. As the columns of the data matrix are usually There are several advantages assumed to be independent, $\mathbf{\Delta} = \mathbf{I}$, this should be our default position. By placing an L_1 penalty on $\mathbf{\Delta}^{-1}$, our model

encourages sparsity in the off-diagonal elements of $\mathbf{\Delta}$. Also, notice that we have one penalty parameter, λ, that is modulated by the dimension of the rows and columns. (We note that the penalty parameter, λ, can be selected by cross-validation). Thus, the evidence of a partial correlation among columns must be strong relative to the correlations among the rows for an off-diagonal element of $\mathbf{\Delta}^{-1}$ to be estimated as non-zero. Secondly,specifically for microarrays, it seems reasonable to assume that the covariances among the genes is sparse, since biologically genes are likely only to be correlated with genes in the same or related pathways.

We also pause briefly to discuss the theoretical rationale for using L_1 penalties, instead of, for example, L_2 penalties. Recall that covariance solutions to the TRCM model with L_2 penalties have eigenvectors that are equal to the left and right singular vectors of the data (Allen and Tibshirani, 2010). Thus, the singular vectors of the data would remain the same when sphering with these estimates. In high-dimensional settings, however, it is well established that eigenvectors of empirical covariances are inconsistent (Johnstone and Lu, 2004), and thus, sphering by the L_2 covariance estimates seems ill-advised. While the consistency of L_1 TRCM estimates has not been established, there are consistency results for multivariate covariance estimation with an L_1 penalty. Rothman et al. (2008) show convergence of the multivariate covariance estimate in the Frobenius norm and more importantly for the correlation estimate in the operator norm which implies convergence of the eigenvectors (El Karoui, 2008). These results reveal some of the possible theoretical advantages of using L_1 penalties to estimate the covariances.

4.2 Sphering Algorithm

Based on the matrix decomposition model, (1), we present a method of de-correlating or sphering the data so that the rows and columns are approximately independent. This sphered data can then be used with standard multiple testing procedures to identify significant row variables. Given a data matrix $\mathbf{X}$ with m rows and n columns, we present our sphering algorithm in Algorithm 1.

Algorithm 1 Sphering Algorithm

1. Estimate row and column means, $\hat{\nu}$ and $\hat{\mu}$ forming $\hat{\mathbf{M}}$, and the signal matrix, $\hat{\mathbf{S}}$.
2. Define the noise, $\mathbf{N} \triangleq \mathbf{X} - \hat{\mathbf{M}} - \hat{\mathbf{S}}$. Estimate row and column covariances of noise, $\hat{\mathbf{\Sigma}}$ and $\hat{\mathbf{\Delta}}$ via TRCM.
3. Sphere the noise: $\tilde{\mathbf{N}} \triangleq \hat{\mathbf{\Sigma}}^{-\frac{1}{2}} \mathbf{N} \hat{\mathbf{\Delta}}^{-\frac{1}{2}}$. Form the sphered data matrix: $\tilde{\mathbf{X}} \triangleq \hat{\mathbf{S}} + \tilde{\mathbf{N}}$.

The Sphering Algorithm simply estimates the means and signal according to the matrix decomposition model (1) and then estimates the correlation structure among the rows and columns in the remaining noise. The TRCM covariance estimates, $\hat{\mathbf{\Sigma}}$ and $\hat{\mathbf{\Delta}}$ are used to de-correlated the noise. Here, $\hat{\mathbf{\Sigma}}^{-1/2}$ is the matrix square root of $\hat{\mathbf{\Sigma}}^{-1}$ and $\hat{\mathbf{\Delta}}^{-1/2}$ of

$\hat{\boldsymbol{\Delta}}$. (We use the symmetric square root defined by the following. Let $\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{P}\Lambda\mathbf{P}^T$ be the eigenvalue decomposition of $\hat{\boldsymbol{\Sigma}}^{-1}$, then the symmetric matrix square root is given by $\hat{\boldsymbol{\Sigma}}^{-1/2} = \mathbf{P}\Lambda^{1/2}\mathbf{P}^T$.) Adding the signal back into this sphered noise, we obtain $\tilde{\mathbf{X}}$ which we call the sphered data. One can use this de-correlated data to find significant row variables.

Now, we investigate some of the theoretical properties of the sphered data for the two-class problem introduced in Section 2.2. The sphered data, $\tilde{\mathbf{X}}$ has the following properties.

Proposition 1 *Let* $\mathbf{X} \sim N_{m,n}(\mathbf{M} + \mathbf{S}, \boldsymbol{\Sigma}, \boldsymbol{\Delta})$ *where* $\mathbf{M} = \nu \mathbf{1}_{(n)}^T + \mathbf{1}_{(m)}\mu^T$ *and* $\mathbf{S} = [\psi_1 \mathbf{1}_{(n_1)}^T \;\; \psi_2 \mathbf{1}_{(n_2)}^T]$ *and let* $\tilde{\mathbf{X}}$ *be the sphered data given by Algorithm 1. Then,*

(i) $\mathrm{E}(\tilde{\mathbf{X}}) = \mathbf{S} = [\psi_1 \mathbf{1}_{(n_1)} \;\; \psi_2 \mathbf{1}_{(n_2)}]$,

(ii) $\tilde{\mathbf{X}} - \hat{\mathbf{S}} \sim N_{m,n}\left(\mathbf{0}, \mathbf{0}, \tilde{\boldsymbol{\Sigma}}, \tilde{\boldsymbol{\Delta}}\right)$,

where $\tilde{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}$ *and* $\tilde{\boldsymbol{\Delta}} = \hat{\boldsymbol{\Delta}}^{-\frac{1}{2}} \boldsymbol{\Delta} \hat{\boldsymbol{\Delta}}^{-\frac{1}{2}}$.

Thus, the class signal remains the same between $\mathbf{X}$ and $\tilde{\mathbf{X}}$, and the covariance structure is all that changes. By sphering the noise, the noise of each row in $\tilde{\mathbf{X}}$ becomes a linear combination of the noise in the other rows.

We now study how sphering the data affects the Z and T statistics from Section 2.2. First, the Z-statistic does not change with sphering. The numerator of both the Z and T statistic, $\bar{x}_1 - \bar{x}_2$ is given by $\hat{\psi}_1 - \hat{\psi}_2$, the components of the estimated signal matrix $\hat{\mathbf{S}}$. The denominator of the T-statistic, namely s_{x_1,x_2}, the estimate of the noise, however, changes with sphering. Recall that in Section 2.2, we discussed how the T-statistic does not have a closed form distribution when there are column correlations. After sphering the data, however, the T-statistic on the sphered data follows a scaled t distribution under certain conditions. This is given by the following result.

Proposition 2 *Let* $\mathbf{X} \sim N_{m,n}(\mathbf{M} + \mathbf{S}, \boldsymbol{\Sigma}, \boldsymbol{\Delta})$ *where* $\mathbf{M} = \nu \mathbf{1}_{(n)}^T + \mathbf{1}_{(m)}\mu^T$ *and* $\mathbf{S} = [\psi_1 \mathbf{1}_{(n_1)}^T \;\; \psi_2 \mathbf{1}_{(n_2)}^T]$. *Let* $\tilde{\mathbf{X}}$ *be the sphered data given by Algorithm 1, and let the statistic* $\tilde{T}_i$ *be the statistic for the* i^{th} *row defined by* (5) *for the data* $\tilde{\mathbf{X}}$. *Then under the null hypothesis* $H_0 : \psi_1 = \psi_2$,

$$\text{if } \tilde{\boldsymbol{\Delta}} = \mathbf{I}, \quad \tilde{T}_i \sim \frac{\tilde{\sigma}_i}{\sigma_i}\sqrt{\frac{\eta}{c_n}}\, t_{(n-2)},$$

where $c_n = \frac{1}{n_1} + \frac{1}{n_2}$, $\eta = \sum_{j=1}^n \left(\frac{1}{n_1}\sum_{i \in \mathcal{C}_1} L_{ij} - \frac{1}{n_2}\sum_{i \in \mathcal{C}_2} L_{ij}\right)^2$ *and* $\mathbf{L}$ *is the matrix square root of* $\boldsymbol{\Delta}$, $\sigma_i = \boldsymbol{\Sigma}_{ii}$ *and* $\tilde{\sigma}_i = \tilde{\boldsymbol{\Sigma}}_{ii}$.

Using our sphering algorithm, we obtain test statistics that follow known distributions when the sphered column covariance, $\tilde{\boldsymbol{\Delta}}$ is the identity. If $\tilde{\boldsymbol{\Delta}}$ is instead a diagonal matrix, then a simple scaling of the columns will give the above result. Notice that if the original

data, $\mathbf{X}$, has no column correlations, $\mathbf{\Delta} = \mathbf{I}$, and $\tilde{\sigma}_i = \sigma_i$, then T and $\tilde{T}$ both follow a t distribution with $n-2$ degrees of freedom. Thus, if the data originally follows the correct theoretical null distribution, then sphering the data does not change its null distribution. Also, if the sphered rows are independent, $\tilde{\mathbf{\Sigma}} = \mathbf{I}$, or approximately independent, then the statistics, T_i are independent or approximately independent. We also note that we can often assume that $\tilde{\sigma}_i = \sigma_i$, thus eliminating that coefficient ratio from the distribution. This is especially a reasonable assumption if the rows are scaled prior to applying the sphering algorithm.

Our results in Proposition 2 hold if the sphered column covariance $\tilde{\mathbf{\Delta}}$ is the identity or diagonal. The TRCM model, however, estimates sparse penalized row and column covariances. These penalized estimates will not capture the full covariances, but will instead estimate the major correlations. Thus, in practice, $\tilde{\mathbf{\Delta}}$ and $\tilde{\mathbf{\Sigma}}$ are not likely to be exactly the identity. We have observed in simulations, however, that $\tilde{\mathbf{\Delta}}$ is often diagonal or nearly diagonal, and thus the theoretical results are appropriate.

When calculating p-values for $\tilde{T}$ based on the distribution given in Proposition 2, we must know the value of η which depends on the original column covariance $\mathbf{\Delta}$. One could estimate this from $\hat{\mathbf{\Delta}}$, but since the TRCM framework estimates penalized covariances, an estimate, $\hat{\eta}$, based on $\hat{\mathbf{\Delta}}$ will underestimate the population η. Hence to obtain the null distribution of the $\tilde{T}$-statistics, we have opted to scale $\tilde{T}$ by the variance of the central portion of the observed distribution of the test statistics. This procedure is outlined in Algorithm 4.2 where $\rho_\alpha(x)$ denotes the α^{th} quantile of x. and $I()$ is the indicator function.

Algorithm 2 Scaling by the central portion of $\tilde{T}$.

---

1. Let the expected proportion of null test statistics be $\hat{\pi}_0 = \hat{m}_0/m$.

2. Estimate the variance of the central portion of sphered test statistics:

$$\hat{\sigma}^2_{\tilde{T}}(\hat{\pi}_0) \triangleq \hat{\text{Var}}\left[\tilde{T}_i \ \text{I}\left(\tilde{T}_i \geq \rho_{((1-\hat{\pi}_0)/2)}(\tilde{T}), \tilde{T}_i \leq \rho_{(1-\hat{\pi}_0/2)}(\tilde{T})\right)\right]$$

3. Define the central-matched $\tilde{T}$-statistics: $\tilde{T}^* \triangleq \tilde{T}\sigma_{t_{(n-2)}}(\hat{\pi}_0)/\hat{\sigma}_{\tilde{T}}(\hat{\pi}_0)$, where $\sigma^2_{t_{(n-2)}}(\hat{\pi}_0)$ is the variance of the central portion of the $t_{(n-2)}$ distribution.

We scale by the central portion of the $\tilde{T}$-statistics so that the statistics can be tested against the $t_{(n-2)}$ distribution. Notice that if all of the test statistics are null and $\pi_0 = 1$ then, central-matching the variances reduces to scaling the $\tilde{T}$-statistics. Since under the assumptions of Proposition 2, only the null $\tilde{T}_i$ follow a scaled t-distribution, we do not want statistics corresponding to non-null tests to contaminate the variance estimates. Thus, we recommend using a conservative estimate of π_0, such as 0.8 or 0.9 for microarrays.

By applying our sphering algorithm, we directly account for correlations among the rows and columns. This results in test statistics that more closely follow both their theoretical

nulls and the theoretical assumptions under which common multiple testing procedures are known to control the false discovery rate.

5 Results

We now evaluate the performance of our sphering algorithm through many simulated examples. First, we compare data pre-processed by sphering to the standard row and column centering method on simulations based on the matrix decomposition model, (1). We use simulations from the matrix-variate normal with the structured covariances from the simulation study in Section 3.2 and also with covariances based on the observed dependencies in real microarray data. Finally, we test the robustness of our method and compare it to other methods for modeling dependencies in Section 5.2. For all simulations, the sphering algorithm was applied with the TRCM penalty parameter λ selected by five-fold cross-validation and with statistics scaled by the central portion using $\pi_0 = 0.8$. (Note that the "standard" pre-processing method refers to row and column centering throughout this section.)

5.1 Simulations: Matrix-variate Model

In all of these simulations, the data, $\mathbf{X}$ is simulated from the matrix decomposition model (1) with $m = 250$ rows and $n = 50$ columns. The first 50 rows are non-null given by $\psi_{1,1:25} = 0.5$, $\psi_{1,26:50} = -0.5$, $\psi_{2,1:25} = -0.5$, $\psi_{2,26:50} = 0.5$ and the last 200 elements of ψ_1 and ψ_2 equal to zero.

In Table 3, we present results on a subset of the simulations from our simulation study in Section 3.2. The remaining simulation study results are given in Appendix A. The results in Table 3 show that de-correlating the data matrix yields improvements in 1) statistical power and 2) estimation of the FDR. We briefly illustrate this by examining a specific example from Table 3.2. Take the simulation with parameters $\mathbf{\Sigma}_1$, $\mathbf{\Delta}_1$ and look at the results with 55 tests rejected. We notice that the true FDP for the data pre-processed by the standard method is 0.111 whereas it is lower, 0.105, on the data that was sphered. This results from a favorable re-ordering of the test statistics that gives a higher statistical power, one minus the true FDP. Next, notice that the FDR estimates for the step-up and permutation-based methods are 0.426 and 0.43 respectively for the un-sphered data. These estimates are overly conservative, as the true FDP is 0.111. After sphering, however, the FDR estimates are 0.124 and 0.125 which are much closer to the true FDP of 0.105. Hence, sphering also improves FDR estimation.

Sphering the data as a pre-processing step to multiple testing procedures has many advantages. First in microarrays, the higher statistical power that results from a re-ordering of test statistics is important to scientists who desire the top ranked genes from one microarray study to translate to the top genes in another study. Also, while sphering leads to improvements in FDR estimation, it is still a slightly conservative estimate, as desired, for the true FDP. As with the simulation study, we find that the empirical null based method of Efron (2007) gives an inconsistent estimate of the FDR as it is both a conservative and

	True FDP	FDR Estimates (Benjamini & Hochberg, 1995)	(Storey & Tibshirani, 2003)	(Efron, 2007)
$\mathbf{\Sigma}_1, \mathbf{\Delta} = \mathbf{I}$				
40 tests	0.0725 (0.033)	0.0723 (0.022)	0.0742 (0.022)	0.257 (0.068)
	0.0333 (0.017)	**0.0458 (0.019)**	**0.0452 (0.019)**	**0.153 (0.075)**
45 tests	0.104 (0.034)	0.103 (0.026)	0.105 (0.026)	0.286 (0.07)
	0.0469 (0.02)	**0.0703 (0.025)**	**0.0705 (0.025)**	**0.173 (0.076)**
50 tests	0.124 (0.033)	0.147 (0.028)	0.149 (0.029)	0.32 (0.069)
	0.0822 (0.02)	**0.104 (0.029)**	**0.105 (0.029)**	**0.207 (0.075)**
55 tests	0.169 (0.025)	0.19 (0.03)	0.193 (0.03)	0.344 (0.066)
	0.141 (0.016)	**0.185 (0.035)**	**0.186 (0.035)**	**0.261 (0.067)**
60 tests	0.222 (0.02)	0.255 (0.036)	0.258 (0.036)	0.376 (0.063)
	0.194 (0.012)	**0.233 (0.038)**	**0.234 (0.038)**	**0.284 (0.067)**
$\mathbf{\Sigma}_1, \mathbf{\Delta}_1$				
40 tests	0.0075 (0.0053)	0.0588 (0.019)	0.0578 (0.019)	0.0152 (0.0053)
	0.00278 (0.0028)	**0.00493 (0.0029)**	**0.00469 (0.0029)**	**0.0071 (0.0049)**
45 tests	0.0222 (0.0099)	0.134 (0.037)	0.133 (0.038)	0.0511 (0.025)
	0.00988 (0.0099)	**0.0157 (0.0071)**	**0.0152 (0.007)**	**0.0171 (0.0086)**
50 tests	0.056 (0.016)	0.245 (0.052)	0.247 (0.053)	0.0858 (0.032)
	0.0222 (0.011)	**0.0438 (0.013)**	**0.0434 (0.013)**	**0.0487 (0.018)**
55 tests	0.111 (0.011)	0.426 (0.04)	0.43 (0.041)	0.139 (0.032)
	0.105 (0.0085)	**0.124 (0.02)**	**0.125 (0.02)**	**0.118 (0.018)**
60 tests	0.178 (0.0083)	0.579 (0.052)	0.585 (0.053)	0.167 (0.028)
	0.172 (0.0039)	**0.199 (0.028)**	**0.201 (0.029)**	**0.146 (0.016)**

Table 3: *A subset of the simulation study results: True false discovery proportions (FDP) and FDR estimates with standard errors are given when a pre-specified number of tests are rejected. Results using the sphering algorithm (in bold) are compared to data that has been row and column centered. All data was simulated under the matrix decomposition model,* (1)*, with parameters given in Section 3.2, and repeated ten times. Two sets of values should be compared: the true FDP with sheering to without sphering, and the FDR estimates compared to the true FDP for both with and without sphering.*

liberal estimate for differing numbers of rejected tests.

We also wish, however, to test the performance of our sphering algorithm on data with dependencies more similar to real microarray data. Thus, we build a second simulation study based upon the empirical covariances of the "Cardio" and the "Leukemia" microarrays. For each of the ten repetitions, we sample 250 genes and 50 arrays at random from each microarray. Let $i \in \mathcal{I}$ and $j \in \mathcal{J}$, be the sampled sets of the genes and arrays respectively, and assume $\mathbf{X}$ is the centered data matrix. We then calculate the empirical covariances, $\mathbf{\Delta}_{MLE} = \sum_{i\in\mathcal{I}}\sum_{i'\in\mathcal{I}} X_i^T X_{i'}/m$, and $\mathbf{\Sigma}_{MLE} = \sum_{j\in\mathcal{J}}\sum_{j'\in\mathcal{J}} X_j X_{j'}^T/n$. The data is simulated from the mean-restricted matrix-variate normal with $\mathbf{X} \sim N([\psi_1 \;\; \psi_2], \mathbf{0}, \mathbf{\Sigma}_{MLE}, \mathbf{\Delta}_{MLE})$. Hence, the simulated data follows the observed covariance of the "Cardio" and "Leukemia" studies. Example images and FDR curves from this simulation are given in Figure 4 as well as the simulation results in Table 4.

The results of the structured covariance study, namely improvements in statistical power and FDR estimation, are confirmed on these microarray-based simulations. There are also

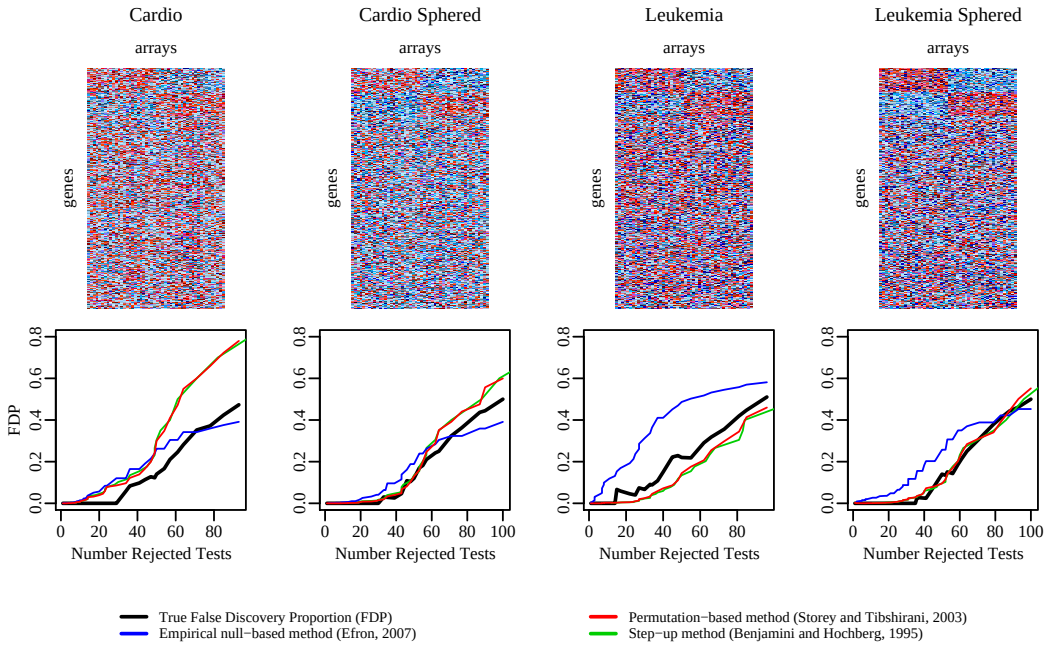

Figure 4: *Example data images (top panel) and FDR curves (bottom panel) for the simulations based on dependencies within the "Cardio" and "Leukemia" microarrays. Data is either gene and array centered or sphered. In the FDR curves, the true and estimated false discovery proportions are plotted against the number of genes rejected. All data was simulated under the matrix decomposition model,* (1)*, with parameters given in Section 5.1.*

some specific notes to make regarding these simulations. First in Table 4, notice that using un-sphered data the FDR is overestimated on the "Cardio" simulations and underestimated on the "Leukemia" simulations. This is confirmed by the example giving the full FDR curves in Figure 4. After sphering, however, we see that the FDR estimates for both simulations are still conservative, but much closer to the true FDP. We note that all ten repetitions of the "Cardio" simulation estimated both $\hat{\mathbf{\Sigma}}$ and $\hat{\mathbf{\Delta}}$ to be non-diagonal. This means that even after accounting for the correlations among the genes, there still appear to be significant correlations among the arrays. In the "Leukemia" simulation, however, $\hat{\mathbf{\Delta}}$ was estimated to be diagonal in all ten simulations. Thus, the correlations among the genes may be driving the over-dispersion seen in the t-statistic distributions of Figure 1.

Thus, from these simulations based on our matrix decomposition model, we see that sphering the data as a pre-processing step greatly improves statistical power and false discovery rate estimation.

	True FDP	FDR Estimates (Benjamini & Hochberg, 1995)	(Storey & Tibshirani, 2003)	(Efron, 2007)
"Cardio"				
40 tests	0.015 (0.01)	0.0794 (0.013)	0.0803 (0.015)	0.0349 (0.016)
	0.00833 (0.0083)	**0.0311 (0.011)**	**0.0301 (0.011)**	**0.0469 (0.021)**
45 tests	0.0333 (0.014)	0.144 (0.019)	0.146 (0.021)	0.0545 (0.021)
	0.0247 (0.013)	**0.0615 (0.02)**	**0.0597 (0.019)**	**0.0757 (0.031)**
50 tests	0.068 (0.017)	0.294 (0.036)	0.295 (0.037)	0.0945 (0.024)
	0.0578 (0.016)	**0.106 (0.026)**	**0.104 (0.025)**	**0.1 (0.029)**
55 tests	0.131 (0.013)	0.452 (0.068)	0.453 (0.068)	0.131 (0.025)
	0.125 (0.015)	**0.196 (0.034)**	**0.195 (0.033)**	**0.153 (0.032)**
60 tests	0.19 (0.011)	0.555 (0.072)	0.555 (0.072)	0.162 (0.022)
	0.191 (0.013)	**0.26 (0.024)**	**0.259 (0.023)**	**0.178 (0.027)**
"Leukemia"				
40 tests	0.0875 (0.016)	0.0777 (0.0068)	0.073 (0.0064)	0.288 (0.053)
	0.0375 (0.012)	**0.0477 (0.013)**	**0.0489 (0.013)**	**0.0928 (0.026)**
45 tests	0.133 (0.023)	0.119 (0.011)	0.112 (0.012)	0.324 (0.054)
	0.0711 (0.016)	**0.0771 (0.017)**	**0.08 (0.018)**	**0.124 (0.031)**
50 tests	0.172 (0.021)	0.162 (0.015)	0.156 (0.015)	0.36 (0.048)
	0.122 (0.017)	**0.129 (0.023)**	**0.132 (0.024)**	**0.159 (0.033)**
55 tests	0.22 (0.019)	0.194 (0.019)	0.186 (0.02)	0.383 (0.046)
	0.178 (0.013)	**0.183 (0.021)**	**0.186 (0.021)**	**0.191 (0.03)**
60 tests	0.255 (0.015)	0.243 (0.024)	0.237 (0.024)	0.405 (0.043)
	0.223 (0.011)	**0.235 (0.022)**	**0.24 (0.023)**	**0.213 (0.03)**

Table 4: *Results for simulations based on observed dependencies within the "Cardio" and "Leukemia" microarrays: True false discovery proportions (FDP) and FDR estimates with standard errors are given when a pre-specified number of tests are rejected. Results using the sphering algorithm (in bold) are compared to data that has been row and array centered. All data was simulated under the matrix decomposition model,* (1)*, with parameters given in Section 5.1, and repeated ten times. Two sets of values should be compared: the true FDP with sheering to without sphering, and the FDR estimates compared to the true FDP for both with and without sphering.*

5.2 Simulations: Other Models

We now evaluate the performance of our method using simulations based on models other than the matrix-variate normal, namely a latent variable model and a random effects model. In these simulations we will not only compare our sphering method to the standard method, but also to the surrogate variable analysis method of Leek and Storey (2008). We first compare this method and model's properties to our sphering algorithm and matrix decomposition model, and then compare these methods numerically.

To account for possible latent variables in a multiple testing framework, Leek and Storey (2008) propose a matrix model and the surrogate variable analysis (SVA) method. They propose the model for the data $\mathbf{X} \in \Re^{m \times n}$, $\mathbf{X} = \mathbf{B}\,\mathbf{S} + \mathbf{\Gamma}\mathbf{G} + \mathbf{U}$ where $\mathbf{S}$ is a signal matrix, $\mathbf{G} \in \Re^{d \times n}$ for $d < n$ is the latent variable matrix, $\mathbf{U} \in \Re^{m \times n}$ is independent noise and $\mathbf{B}$ and $\mathbf{\Gamma} \in \Re^{m \times d}$ are coefficients to be estimated. This model is similar in nature to our matrix decomposition model (1). If we assume $\mathbf{X}$ has been previously centered, using the notation

	Standard		Sphered		SVA	
	FDP	$\widehat{\text{FDR}}$	FDP	$\widehat{\text{FDR}}$	FDP	$\widehat{\text{FDR}}$
Latent Variable Model						
40 tests	0.08	0.168	0.05	0.0365	0.0528	0.0546
	(0.017)	(0.023)	(0.014)	(0.0086)	(0.01)	(0.0045)
45 tests	0.109	0.212	0.0889	0.0657	0.0716	0.0868
	(0.016)	(0.022)	(0.018)	(0.013)	(0.0097)	(0.0092)
50 tests	0.15	0.303	0.124	0.106	0.127	0.118
	(0.018)	(0.035)	(0.017)	(0.021)	(0.01)	(0.015)
55 tests	0.189	0.383	0.167	0.166	0.17	0.183
	(0.015)	(0.051)	(0.018)	(0.021)	(0.012)	(0.02)
60 tests	0.24	0.453	0.215	0.215	0.217	0.27
	(0.011)	(0.05)	(0.017)	(0.023)	(0.0099)	(0.025)
Random Effects Model						
40 tests	0.375	0.011	0.0361	0.0318	0.4	0.136
	(0.011)	(0.002)	(0.02)	(0.012)	(0.047)	(0.042)
45 tests	0.433	0.0161	0.0642	0.0863	0.439	0.169
	(0.0084)	(0.0031)	(0.026)	(0.031)	(0.042)	(0.048)
50 tests	0.48	0.0196	0.102	0.14	0.475	0.185
	(0.014)	(0.004)	(0.02)	(0.033)	(0.041)	(0.047)
55 tests	0.52	0.0229	0.154	0.207	0.507	0.213
	(0.013)	(0.0044)	(0.018)	(0.037)	(0.034)	(0.053)
60 tests	0.553	0.0288	0.202	0.309	0.538	0.235
	(0.012)	(0.0054)	(0.014)	(0.048)	(0.031)	(0.053)

Table 5: *Simulations based on a latent variable and random-effects models as described in Section 5.2. The true FDP (FDP) is compared to the FDR estimates ($\widehat{\text{FDR}}$) for a fixed number of rejected tests using the step-up method of Benjamini and Hochberg (1995) for three pre-processing techniques: Standard (row and column centered), our sphering algorithm, and the surrogate variable analysis (SVA) method. Averages are taken on ten repetitions and standard errors are given.*

of the latent variable model, we can write (1) as $\mathbf{X} = \mathbf{B}\,\mathbf{S} + \mathbf{\Sigma}^{\frac{1}{2}}\,\mathbf{U}\,\mathbf{\Delta}^{\frac{1}{2}}$. Thus, our model accounts for structure within the data through the row and column covariances $\mathbf{\Sigma}$ and $\mathbf{\Delta}$, whereas their method estimates the structure through $\mathbf{G}$ and assumes the noise is additive. Assuming that the latent variable model or our model is correct, applying the respective algorithms results in approximately independent p-values. Also similar to our method, SVA can change the rankings of the test statistics. Unlike SVA, however, our model and sphering algorithm directly capture and account for possible correlations among the columns as well as the rows.

We simulate data from a latent variable model taken directly from Leek and Storey (2008) as well as from a random effects model denoting a batch effect. For both models, the data is of dimension 250×50, with 25 columns in each class with the following signal: The first 50 rows are non-null given by $\psi_{1,1:25} = 0.5$, $\psi_{1,26:50} = -0.5$, $\psi_{2,1:25} = -0.5$, $\psi_{2,26:50} = 0.5$ and the last 200 elements of ψ_1 and ψ_2 equal to zero. For the latent variable model, there are two latent variables given by $G_{ij} \overset{iid}{\sim} Bern(0.5)$, coefficients $\Gamma_{ij} \overset{iid}{\sim} N(0,1)$ and noise $U_{ij} \overset{idd}{\sim} N(0,1)$. For the random-effects model with K batches indicated by indices

$I(k)$, a column of the data is given by the following: $X_{rj} = \nu + \mu_j + \sum_{i=1}^{2} \psi_i I_{(j \in \mathcal{C}_i)} + \sum_{k=1}^{K} \beta_k I_{(j \in I(k))} + \epsilon$, where ν, μ_j and ψ_i are fixed effects, and $\beta_k \overset{iid}{\sim} N(\mu_k, \sigma_k^2 \mathbf{I})$ independent of $\epsilon \overset{iid}{\sim} N(0, \mathbf{\Sigma}_1)$ are random effects. In our simulation, we have $\mu_k = [-0.5 \ \ -0.25 \ \ 0 \ \ 0.25 \ \ 0.5]$, $\sigma^2 = 0.5$, $\mathbf{\Sigma} = \mathbf{\Sigma}_1$ as defined in Section 3.2 and $I(k)$ indicating batches of five columns.

In Table 5, we compare the true FDP to the estimate of the FDR via the step-up method (Benjamini and Hochberg, 1995) for the data with standard pre-processing, with sphering and with Leek and Storey (2008)'s surrogate variable analysis (SVA). The SVA method was implemented using the defaults available in the package sva from CRAN, the R language repository. For the latent variable simulation, both our sphering method and the SVA method improve the rank ordering of the test statistics resulting in higher statistical power as well as improved estimates of the FDR. In the random-effects model simulation, however, the sphering algorithm substantially outperforms the standard pre-processing and the SVA method. We illustrate this by looking at the specific case where 50 tests are rejected. For the standard pre-processing and SVA methods, the true FDP is 0.48 and 0.475 respectively, meaning that on average 25 out of the 50 rejected tests are false positives. With sphering, however, the order of the test statistics is dramatically changed leading to a true FDP of 0.102 so that on average only 5 out of 50 rejected tests are false. The FDR estimates using the step-up method are also problematic for the standard and SVA methods as 0.0196 and 0.185 are substantially below the true FDPs of 0.48 and 0.475 respectively. If these methods were used, the number of false positives would not be controlled. With sphering, however, the FDR estimate of 0.14 is much closer to the true FDP, 0.102 and is a conservative estimate, as desired.

These simulations based on models other than the matrix-variate normal reveal the robustness of our pre-processing technique. Our sphering method also compares very favorably to the surrogate variable analysis, another pre-processing method.

6 Discussion

In this paper, we have demonstrated that using standard statistical methodology to conduct inference on transposable data is problematic. As a method of solving these problems, we have prosed a sphering pre-processing technique that de-correlates the data yielding approximately independent rows and columns. We have revealed the advantages and robustness of this method through simulations on many correlated data sets.

The major disadvantage of our method is its computational cost. Fitting the transposable regularized covariance model with L_1 penalties is approximately $O(k(m^3+p^3))$, where k is the total number of iterations needed until convergence. Thus, directly fitting this model to microarrays, for example, where m may be twenty or thirty thousand, is not currently feasible. A simple fix can be proposed, that is to first filter the genes by the absolute value of their un-sphered T-statistics down to say 1,000 or 500 genes. Since the signal in each gene remains the same before and after sphering, filtering should not effect the power to detect non-null genes, especially since researchers are rarely interested in re-testing over 500

genes. As future work, we will examine approximations to the TRCM covariance estimates that can be used in high-dimensional settings and would circumvent the need to filter the genes before sphering.

There are many components of our work that deserve further investigation and testing. First, Allen and Tibshirani (2010) outline some of the properties of the TRCM covariances estimates, but several questions, such as the consistency of the estimates, remain. Also, direct estimation of η, the scaled variance of the Z-statistic that depends on the array covariance, should be examined to find a consistent estimate of η.

In conclusion, our model and study have revealed several important issues related to large-scale inference with transposable data such as microarrays. First, correlation among the columns proves to be a major problem, both theoretically and in simulations, when comparing test statistics to a theoretical or permutation null distribution. This results is striking as inference is often conducted under the false assumption of column independence. Second, despite the lack of theoretical results supporting the use of many common FDR-controlling procedures for test statistics with arbitrary dependence structure, the procedures seem to conservatively estimate the FDP under a variety of correlation scenarios. Finally, our method of de-correlating the data is a way to directly model the covariance structure in a multiple testing framework. This method leads to 1) improvements in the statistical power, and to 2) better estimation of the FDR. While this paper has focused on the example of two-class microarrays, our model and methods may prove useful in a variety large-scale inference problems with highly transposable data sets.

7 Acknowledgments

We would like to thank Jonathan Taylor for several helpful comments and conversations regarding this work. Thanks to Joseph Romano for discussions and references for papers on multiple testing with dependencies. Thanks also to Bradley Efron whose observations and ideas on microarrays partly inspired this work.

A Additional Simulation Results

B Proofs

Proof 1 (Theorem 1) *Let* $\mathbf{z}$ *be a vector of* $N(0,1)$ *random variables. Then, if we arrange* $\mathbf{x}$ *as a column vector, we have*

$$\mathbf{x}_{(n)} \stackrel{d}{=} \begin{pmatrix} \psi_1 \mathbf{1}_{(n_1)} \\ \psi_2 \mathbf{1}_{(n_2)} \end{pmatrix} + \sigma \mathbf{L} \mathbf{z}_{(n)}.$$

	True FDP	FDR Estimates (Benjamini & Hochberg, 1995)	(Storey & Tibshirani, 2003)	(Efron, 2007)
$\mathbf{\Sigma}_1, \mathbf{\Delta}_2$				
40 tests	0.0125 (0.0077)	0.0831 (0.018)	0.0783 (0.019)	0.0749 (0.024)
	0.0194 (0.013)	**0.022 (0.0074)**	**0.0215 (0.0072)**	**0.0495 (0.029)**
45 tests	0.0333 (0.015)	0.164 (0.031)	0.16 (0.032)	0.117 (0.032)
	0.0321 (0.018)	**0.0436 (0.011)**	**0.0431 (0.011)**	**0.0761 (0.03)**
50 tests	0.078 (0.015)	0.281 (0.027)	0.276 (0.028)	0.165 (0.032)
	0.0556 (0.02)	**0.0883 (0.013)**	**0.0875 (0.013)**	**0.117 (0.031)**
55 tests	0.135 (0.012)	0.368 (0.028)	0.364 (0.028)	0.194 (0.028)
	0.123 (0.015)	**0.181 (0.02)**	**0.182 (0.02)**	**0.179 (0.031)**
60 tests	0.198 (0.011)	0.461 (0.044)	0.458 (0.045)	0.234 (0.028)
	0.194 (0.014)	**0.242 (0.023)**	**0.244 (0.023)**	**0.201 (0.033)**
$\mathbf{\Sigma}_2, \mathbf{\Delta} = \mathbf{I}$				
40 tests	0.035 (0.017)	0.0498 (0.0081)	0.0513 (0.0082)	0.161 (0.034)
	0.03 (0.015)	**0.0279 (0.0058)**	**0.0279 (0.0055)**	**0.081 (0.029)**
45 tests	0.0711 (0.019)	0.0839 (0.012)	0.0856 (0.012)	0.209 (0.041)
	0.0644 (0.017)	**0.0607 (0.0096)**	**0.0604 (0.0094)**	**0.133 (0.035)**
50 tests	0.12 (0.018)	0.141 (0.021)	0.143 (0.021)	0.249 (0.045)
	0.098 (0.013)	**0.0989 (0.013)**	**0.0985 (0.013)**	**0.161 (0.035)**
55 tests	0.167 (0.016)	0.191 (0.029)	0.192 (0.029)	0.278 (0.04)
	0.14 (0.009)	**0.158 (0.016)**	**0.16 (0.016)**	**0.208 (0.035)**
60 tests	0.217 (0.014)	0.243 (0.035)	0.246 (0.035)	0.311 (0.041)
	0.197 (0.0065)	**0.227 (0.017)**	**0.229 (0.017)**	**0.242 (0.034)**
$\mathbf{\Sigma}_2, \mathbf{\Delta}_1$				
40 tests	0.005 (0.005)	0.0455 (0.0065)	0.0439 (0.0065)	0.00846 (0.0041)
	0 (0)	**0.00305 (0.0014)**	**0.00267 (0.0012)**	**0.00185 (0.001)**
45 tests	0.0111 (0.005)	0.111 (0.027)	0.11 (0.026)	0.0198 (0.0076)
	0 (0)	**0.00845 (0.0031)**	**0.00783 (0.0029)**	**0.00656 (0.0026)**
50 tests	0.042 (0.0081)	0.225 (0.046)	0.225 (0.047)	0.0493 (0.014)
	0.03 (0.0061)	**0.0436 (0.0076)**	**0.0433 (0.0075)**	**0.033 (0.0083)**
55 tests	0.109 (0.0086)	0.404 (0.034)	0.409 (0.034)	0.0923 (0.017)
	0.0964 (0.0039)	**0.118 (0.014)**	**0.118 (0.015)**	**0.0756 (0.014)**
60 tests	0.178 (0.0056)	0.552 (0.048)	0.554 (0.048)	0.133 (0.018)
	0.168 (0.0017)	**0.214 (0.014)**	**0.216 (0.014)**	**0.114 (0.015)**
$\mathbf{\Sigma}_2, \mathbf{\Delta}_2$				
40 tests	0.0075 (0.0053)	0.0712 (0.016)	0.066 (0.015)	0.0444 (0.012)
	0.0125 (0.0077)	**0.0108 (0.0021)**	**0.0104 (0.0019)**	**0.0152 (0.0042)**
45 tests	0.0311 (0.011)	0.169 (0.022)	0.163 (0.021)	0.087 (0.019)
	0.0244 (0.0084)	**0.0317 (0.0056)**	**0.0309 (0.0057)**	**0.0336 (0.007)**
50 tests	0.078 (0.012)	0.277 (0.025)	0.271 (0.025)	0.132 (0.021)
	0.072 (0.015)	**0.092 (0.016)**	**0.0918 (0.016)**	**0.0738 (0.011)**
55 tests	0.144 (0.013)	0.388 (0.037)	0.383 (0.037)	0.167 (0.019)
	0.136 (0.013)	**0.162 (0.013)**	**0.162 (0.013)**	**0.116 (0.013)**
60 tests	0.198 (0.013)	0.47 (0.044)	0.468 (0.043)	0.197 (0.02)
	0.202 (0.012)	**0.223 (0.014)**	**0.223 (0.014)**	**0.143 (0.01)**

Table 6: *Additional simulation study results: True false discovery proportions (FDP) and FDR estimates with standard errors are given when a pre-specified number of tests are rejected. Results using the sphering algorithm (in bold) are compared to data that has been row and column centered. All data was simulated under the matrix decomposition model,* (1)*, with parameters given in Section 3.2, and repeated ten times. Two sets of values should be compared: the true FDP with sphering to without sphering, and the FDR estimates compared to the true FDP for both with and without sphering.*

Thus, we can write Z as a sum of the scaled independent and normally distributed random variables $\mathbf{z}$. The expected value of Z is trivial and the variance can be written as the following.

$$\begin{aligned}
\mathrm{Var}(Z) &= \frac{1}{\sigma^2 c_n}\mathrm{Var}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right) \\
&= \frac{1}{c_n}\mathrm{Var}\left(\frac{1}{n_1}\sum_{i\in\mathcal{C}_1}(\mathbf{Lz})_i - \frac{1}{n_2}\sum_{i\in\mathcal{C}_2}(\mathbf{Lz})_i\right) \\
&= \frac{1}{c_n}\mathrm{Var}\left[\sum_{j=1}^{n}\left(\frac{1}{n_1}\sum_{i\in\mathcal{C}_1}L_{ij}z_j - \frac{1}{n_2}\sum_{i\in\mathcal{C}_2}L_{ij}z_j\right)\right] \\
&= \frac{\eta}{c_n}
\end{aligned}$$

The last step follows since the z_j's are independent. Note also that if we let $W_i \triangleq \begin{cases} \frac{1}{n_1} & i \in \mathcal{C}_1 \\ -\frac{1}{n_2} & i \in \mathcal{C}_2 \end{cases}$, we can write $\eta = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{\Delta}_{ij}W_iW_j$.

Proof 2 (Corollary 1) *This is trivial following the proof of Theorem 1 since the matrix square root of $\mathbf{\Delta}$ can be written as $\mathbf{L} = \begin{pmatrix} \mathbf{L}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_2 \end{pmatrix}$.*

Proof 3 (Proposition 1) *For part (i), the matrix decomposition model implies that $\mathrm{E}(X_{ij}) = \nu_i + \mu_j + \psi_{k,i}$ for $k = 1, 2$ depending on the class of array j. Each element of the noise as defined in Step 2 can be written as $N_{ij} = X_{ij} - \hat{\mu}_j - \hat{\nu}_i - \hat{\psi}_{k,i}$ where $\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n}X_{ij}$, $\hat{\nu}_i = \frac{1}{p}\sum_{j=1}^{p}(X_{ij} - \hat{\mu}_j)$, and $\hat{\psi}_{k,i} = \frac{1}{n_k}\sum_{j\in\mathcal{C}_k}(X_{ij} - \hat{\mu}_j - \hat{\nu}_i)$. We show that $\mathrm{E}(N_{ij}) = 0$, which in the process proves that $\mathrm{E}(\tilde{X}_{ij}) = \psi_{k,i}$.*

$$\begin{aligned}
\mathrm{E}(\hat{\mu}_j + \hat{\nu}_i + \hat{\psi}_{k,i}) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_{ij}) + \frac{1}{n_k}\sum_{j\in\mathcal{C}_k}\mathrm{E}(X_{ij}) - \frac{1}{n}\frac{1}{n_k}\sum_{j\in\mathcal{C}_k}\sum_{i=1}^{n}\mathrm{E}(X_{ij}) \\
&= \mu_j + \bar{\nu} + \bar{\psi}_k + \bar{\mu}_k + \nu_i - \psi_{k,i} - \bar{\mu}_k - \bar{\nu} - \bar{\psi}_k \\
&= \mu_j + \nu_i + \psi_{k,i}.
\end{aligned}$$

Now for part (ii), $\tilde{\mathbf{X}} - \mathbf{S} = \tilde{\mathbf{N}} \triangleq \hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\,\mathbf{N}\,\hat{\mathbf{\Delta}}^{-\frac{1}{2}}$, and following the proof of part (i), $\mathbf{N} \sim N_{m,n}(\mathbf{0}, \mathbf{0}, \mathbf{\Sigma}, \mathbf{\Delta})$. The characteristic function of the centered matrix-variate normal is $\phi_{\mathbf{X}}(\mathbf{Z}) = \mathrm{etr}(-\frac{1}{2}\,\mathbf{Z}^T\,\mathbf{\Sigma}\,\mathbf{Z}\,\mathbf{\Delta})$ where etr is the exponential of the trace function (Gupta and Nagar, 1999). The characteristic function of $\tilde{\mathbf{N}}$ can then be written as

$$\begin{aligned}
\phi_{\tilde{\mathbf{N}}}(\mathbf{Z}) &= \mathrm{etr}\left[-\frac{1}{2}\left(\hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\,\mathbf{Z}\,\hat{\mathbf{\Delta}}^{-\frac{1}{2}}\right)^T\mathbf{\Sigma}\left(\hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\,\mathbf{Z}\,\hat{\mathbf{\Delta}}^{-\frac{1}{2}}\right)\mathbf{\Delta}\right] \\
&= \mathrm{etr}\left[-\frac{1}{2}\hat{\mathbf{\Delta}}^{-\frac{1}{2}}\,\mathbf{Z}^T\,\hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\,\mathbf{\Sigma}\,\hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\,\mathbf{Z}\,\hat{\mathbf{\Delta}}^{-\frac{1}{2}}\,\mathbf{\Delta}\right] \\
&= \mathrm{etr}\left[-\frac{1}{2}\,\mathbf{Z}^T\left(\hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\,\mathbf{\Sigma}\,\hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\right)\mathbf{Z}\left(\hat{\mathbf{\Delta}}^{-\frac{1}{2}}\,\mathbf{\Delta}\,\hat{\mathbf{\Delta}}^{-\frac{1}{2}}\right)\right].
\end{aligned}$$

Thus, letting $\tilde{\mathbf{\Sigma}} = \hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\,\mathbf{\Sigma}\,\hat{\mathbf{\Sigma}}^{-\frac{1}{2}}$ and $\tilde{\mathbf{\Delta}} = \hat{\mathbf{\Delta}}^{-\frac{1}{2}}\,\mathbf{\Delta}\,\hat{\mathbf{\Delta}}^{-\frac{1}{2}}$, we have $\tilde{\mathbf{N}} \sim N_{m,n}(\mathbf{0}, \mathbf{0}, \tilde{\mathbf{\Sigma}}, \tilde{\mathbf{\Delta}})$.

Proof 4 (Proposition 2) *Since we are considering the test statistic for one gene, we will suppress the index i. We can define the random variables $Z \triangleq (\bar{X}_1 - \bar{X}_2)/\sigma\sqrt{c_n}$ and $D \triangleq s^2_{\tilde{X}_1\tilde{X}_2}/\sigma^2$. Then, $\tilde{T}$*

can be written as $\tilde{T} = Z/\sqrt{D}$. *From Theorem 1,* $Z \sim N((\psi_1 - \psi_2)/\sigma\sqrt{c_n}, \eta/c_n)$. *Then, under the null,* $Z/\sqrt{\eta/c_n} \sim N(0,1)$. *Also, under the null,* $D\sigma^2/\tilde{\sigma}^2 \sim \chi^2_{(n-2)}$ *with* D *and* Z *independent. Then,*

$$\frac{Z/\sqrt{\frac{\eta}{c_n}}}{\sqrt{D\frac{\sigma^2}{\tilde{\sigma}^2}}} \sim t_{(n-2)} \quad \Rightarrow \quad \tilde{T} = \frac{Z}{\sqrt{D}} \sim \frac{\tilde{\sigma}}{\sigma}\sqrt{\frac{\eta}{c_n}}t_{(n-2)}.$$

References

Allen, G. I. and R. Tibshirani (2010). Transposable regularized covariance models with an applicaiton to missing data imputation. *Ann. Appl. Statist.* –(–), –.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological) 57*(1), 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics 29*, 1165–1188.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association 99*, 96–104.

Efron, B. (2007). Size, power and false discovery rates. *Ann. Statist. 35*(4), 1351–1377.

Efron, B. (2009). Are a set of microarrays independent of each other? *Ann. App. Statist. 13*(3), 922–942.

El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist. 36*(6), 2717–2756.

Ge, Y., S. Dudoit, and T. P. Speed (2003). Resampling-based multiple testing for microarray data analysis. *TEST 12*, 1–77.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science 286*(5439), 531–537.

Gupta, A. K. and D. K. Nagar (1999). *Matrix variate distributions*, Volume 104 of *Monographs and Surveys in Pure and Applied Mathematics*. Boca Raton, FL: Chapman & Hall, CRC Press.

Hommel, G. (1986). Multiple test procedures for arbitrary dependence structures. *Metrika 33*(1), 321–336.

Johnstone, I. M. and A. Y. Lu (2004). Sparse principal components analysis. Unpublished manuscript.

Leek, J. T. and J. D. Storey (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences 105*(48), 18718–18723.

Olshen, R. A. and B. Rajaratnam (2010). Successive normalization of rectangular arrays. *Ann. Statist. 38*(3), 1638–1664.

Owen, A. B. (2005). Variance of the number of false discoveries. *Journal Of The Royal Statistical Society Series B 67*(3), 411–426.

Qiu, X., A. I. Brooks, L. Klebanov, and A. Yakovlev (2005). The effects of normalization on the correlation structure of microarray data. *BMC Bioinformatics 6*(120), –.

Rothman, A. J., P. J. Bickel, E. Levina, and J. Zhu (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat. 2*, 494–515.

Sarkar, S. K. (2008). On methods controlling the false discovery rate. *Sankhya : The Indian Journal of Statistics 70*, 135–168.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal Of The Royal Statistical Society Series B 64*(3), 479–498.

Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal Of The Royal Statistical Society Series B 66*(1), 187–205.

Storey, J. D. and R. Tibshirani (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences 100*(16), 9440–9445.

Tusher, V. G., R. Tibshirani, and G. Chu (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America 98*(9), 5116–5121.

Yekutieli, D. and Y. Benjamini (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference 82*(1-2), 171 – 196.